International Journal of
Molecular Sciences

MDPI

Article

Molecular Modelling Hurdle in the Next-Generation Sequencing Era

Guerau Fernandez [1,2], Dèlia Yubero [1,2,*], Francesc Palau [1,2,3] and Judith Armstrong [1,2]

[1] Department of Genetic and Molecular Medicine—IPER, Hospital Sant Joan de Déu, Institut de Recerca Sant Joan de Déu, 08950 Barcelona, Spain; guerau.fernandez@sjd.es (G.F.); francesc.palau@sjd.es (F.P.); judith.armstrong@sjd.es (J.A.)
[2] Center for Biomedical Research Network on Rare Diseases (CIBERER), ISCIII, 08950 Barcelona, Spain
[3] Division of Pediatrics, University of Barcelona School of Medicine and Health Sciences, 08007 Barcelona, Spain
* Correspondence: delia.yubero@sjd.es; Tel.: +34-93-600-9451; Fax: +34-93-600-9760

Abstract: There are challenges in the genetic diagnosis of rare diseases, and pursuing an optimal strategy to identify the cause of the disease is one of the main objectives of any clinical genomics unit. A range of techniques are currently used to characterize the genomic variability within the human genome to detect causative variants of specific disorders. With the introduction of next-generation sequencing (NGS) in the clinical setting, geneticists can study single-nucleotide variants (SNVs) throughout the entire exome/genome. In turn, the number of variants to be evaluated per patient has increased significantly, and more information has to be processed and analyzed to determine a proper diagnosis. Roughly 50% of patients with a Mendelian genetic disorder are diagnosed using NGS, but a fair number of patients still suffer a diagnostic odyssey. Due to the inherent diversity of the human population, as more exomes or genomes are sequenced, variants of uncertain significance (VUSs) will increase exponentially. Thus, assigning relevance to a VUS (non-synonymous as well as synonymous) in an undiagnosed patient becomes crucial to assess the proper diagnosis. Multiple algorithms have been used to predict how a specific mutation might affect the protein's function, but they are far from accurate enough to be conclusive. In this work, we highlight the difficulties of genomic variability determined by NGS that have arisen in diagnosing rare genetic diseases, and how molecular modelling has to be a key component to elucidate the relevance of a specific mutation in the protein's loss of function or malfunction. We suggest that the creation of a multi-omics data model should improve the classification of pathogenicity for a significant amount of the detected genomic variability. Moreover, we argue how it should be incorporated systematically in the process of variant evaluation to be useful in the clinical setting and the diagnostic pipeline.

Keywords: NGS; VUS; multi-omics; tissue-specific model

Citation: Fernandez, G.; Yubero, D.; Palau, F.; Armstrong, J. Molecular Modelling Hurdle in the Next-Generation Sequencing Era. *Int. J. Mol. Sci.* **2022**, *23*, 7176. https://doi.org/10.3390/ijms23137176

Academic Editor: Ray Luo

Received: 9 June 2022
Accepted: 27 June 2022
Published: 28 June 2022

1. Introduction

Since the release of the first draft of the human genome in 2001 [1], one of the most fundamental challenges clinical geneticists have faced is trying to uncover the cause of Mendelian or single-gene disorders. The lack of functional knowledge of most of the variability within the genome has been the main barrier for genetic diagnostics. In an effort to understand the genetic complexity, rare diseases (RDs) have been the focus of active research. Approximately 72% of RDs are caused by genetic mutations. Due to the singularity of RDs, a disease or condition is considered rare if it affects less than 1 in 200,000 people (United States) or less than 1 in 2000 people (Europe) [2]. Unexplored regions of the genome unlock its function, leading to a distinguishable phenotype. The identification of one or multiple variants that trigger an RD maps a genomic localization to a function (genotype–phenotype). The obvious handicap of working with RDs is the difficulty in establishing a cause–effect association of a specific mutation, requiring functional studies.

Although mutation hotspots have been described for specific RDs, most disease-causing genes have pathogenic mutations scattered all over the gene body or regulatory regions. Identifying the same mutation in different patients with the same pathophysiological processes is difficult in the RD context. To solve this conundrum, two strategies have been applied: The first approach involves investigation of familiar or highly consanguineous populations to increase the probability of characterizing recurrent mutations detected throughout a well-described pedigree, which correlates with the phenotype of study. The second procedure requires gathering and aggregating patients with the same phenotype on a global scale. An example of this kind of initiative is Matchmaker Exchange [3], where genotypic and phenotypic data are shared, seeking common variability to reveal the putative etiology of an undiagnosed group of patients. This second procedure provides less sensitive results compared to the first strategy, as genetic alterations among different patients might not be the same. Overall, 9349 clinical entities and 7877 rare-disease-gene linkages have been described [4].

Since 2009, when next-generation sequencing (NGS) was first introduced to identify disease-causing genes [5], the pace of RD-causing gene discovery has increased drastically [6]. Due to its high resolution, being able to identify mutations at the base-pair level, and its capability to analyze multiple genes at the same time, NGS has become one of the most powerful tools to detect genetic variability. Thus, NGS has been fully incorporated into the clinical setting for disease diagnosis, in combination with other, more conventional techniques [7]. The use of NGS has allowed a deep characterization and subgrouping of certain diseases, leading to more accurate diagnosis. However, as a counterbalance to this precise genetic profiling, phenotypic overlap in human diseases has increased since the discovery of new causative genes, blurring the lines between diseases.

NGS has driven genetic diagnosis of RDs to a global 50% of diagnostic yield. Some diseases present diagnostic yields over 70%—for example, inborn errors of metabolism or specific neurological conditions, in which the presence of a biomarker facilitates the genetic diagnosis [8,9]. Moreover, having biological support for a genetic disorder allows the possibility to design other strategies to resolve unsolved conditions, such as looking for intronic or regulatory regions, or using specific approaches to detect structural variation. Unfortunately, the diagnostic odyssey remains harder for patients with RDs that do not fall into these groups, and understanding their genomes becomes a complex process. It is worth mentioning the reusability of NGS data. Variants of uncertain significance (VUSs) detected in an undiagnosed patient can be reclassified as disease-causing in light of new discoveries.

Alongside the significant increase in diagnostic yield, NGS has boosted the amount of data generated and, therefore, has increased uncertainty. The more genomes are sequenced, the more VUSs are obtained. Moreover, due to pleiotropic effects, other factors are putting molecular analysis alone into a deadlock; a single gene could affect multiple and apparently unrelated phenotypes and mutational penetrance, and a particular mutation does not always produce the same effect in all individuals who carry it (Figure 1).

The aim of this work is to portray the actual state of molecular diagnosis, considering only genomic sequencing, and to study complementary sources of information in order to combine them and determine a model to make better use of these powerful data. Omics data have to be structured and integrated in order to achieve a clearer picture of the possible impact that novel and unknown variants might have. It is imperative for the benefit of patients with RDs that we drive diagnostic yield to as close to 100% as possible for as many conditions as possible [10], using all of the tools we have available.

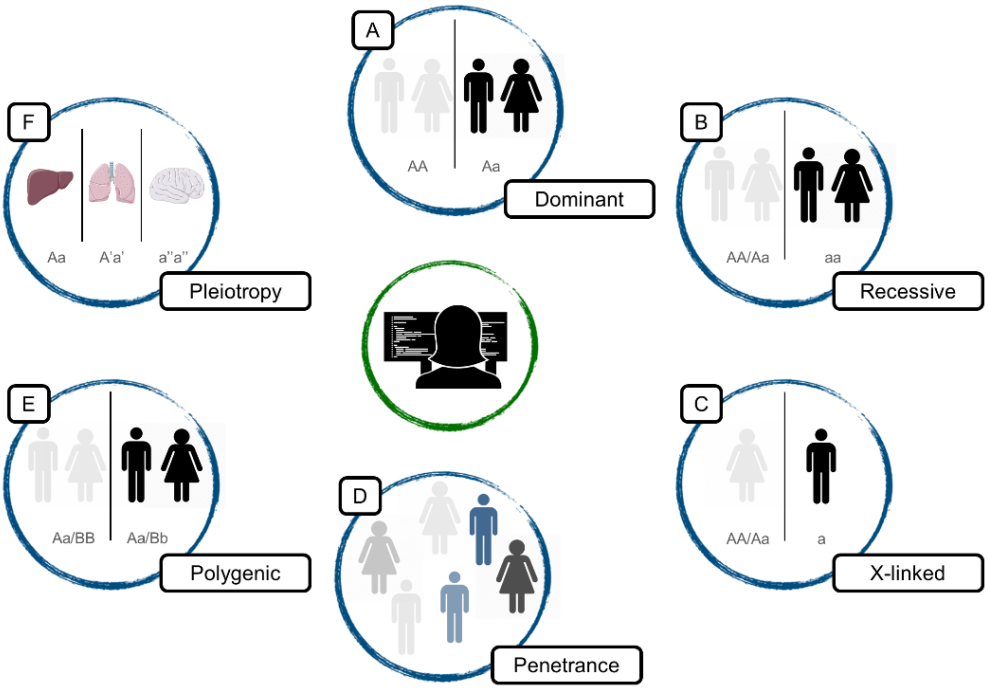


Figure 1. Examples of gene effect complexity. (**A**). Dominant: a single mutated allele is enough to cause the disorder (**B**). Recessive: two mutated alleles are required to cause the disorder (**C**). X-linked: males that carries the disease-causing mutation are affected due to the single copy of the X chromosome (**D**). Penetrance: same genetic variant might not develop the same symptomatology in different individuals (**E**). Polygenic: disorder caused by the combined action of more than one gene (**F**). Pleiotropy: mutations in a single gene affects two or more apparently unrelated disorders.

2. Results

2.1. Genetic Variability

Considering the fraction of DNA sequenced by NGS technology, genomic variation can be determined using gene panels, which can span from less than a hundred to a few thousand protein-coding genes; whole-exome sequencing (WES), which comprises genes with well-established pathogenicity as well as functionally unknown genes; and whole-genome sequencing (WGS), where variants in the intergenic non-coding regions can be analyzed. Panels of all known pathogenic genes—also known as clinical exome sequencing (CES)—are one of the choices to introduce NGS into the clinical setting.

To illustrate the limitations that a regular genetics department encounters routinely when analyzing NGS data, we studied the outcomes derived from the Illumina TruSight One Expanded (TSOex) gene panel (6704 genes) from all patients analyzed over 2 years (n = 2474) at the Sant Joan de Déu Children's Hospital. Genes included in TSOex are abundant in cellular processes, related to binding or catalytic activity, mostly belong to the metabolite interconversion enzyme protein class, and are enriched in 71 OMIM (Online Mendelian Inheritance in Man) disease categories, with anemia being the most significant (Figure 2A–D and Supplementary Table S1).

As genetic variants with low population frequency are more likely to be the cause of genetic diseases, for this study, we selected variants with an allele frequency < 0.01 (European non-Finnish, gnomAD). A total of 2,456,984 variants were detected within all samples. The five most abundant types of variants per sample are shown in Figure 3A; missense mutations were the most prominent. Although synonymous mutations are also very well represented, when determining the unique mutations within all samples, their proportion becomes almost negligible (Figure 3B). The drastic reduction in recurrent synonymous mutations might be the result of common variants in the Southern European population that are underrepresented in the gnomAD database. More than 90% of the

mutations are missense, followed by stop-gain and structural interaction mutations. Next, we determined the number of mutations detected per gene, and summarized all samples (Figure 3C). Most genes only showed one mutation. TTN was the most mutated gene, with 2337 individual variants throughout its gene body. Interestingly, there was a slight peak between 8 and 50 mutations per gene, followed by a long tail of increasingly mutated genes. Due to the differential mutational landscape among the analyzed genes, we compared them to their lack of permissiveness to accept missense variations that might lead to loss of protein function, using the observed versus expected values from gnomAD (Figure 3D). The higher (closer to 0) this score, the more tolerant the gene, and the lower this score (negative values), the more intolerant it is. Most genes are tolerant to variations, showing two high-density areas: one at 1 mutation per gene (for example, the case of gene RPS17), and another at close to 50 mutations per gene (for example, the gene LRPAP1). Surprisingly, the most intolerant genes correlate with the 8–50 peak shown in Figure 3C, indicating that mutations in those genes are the most susceptible to leading to loss of function.

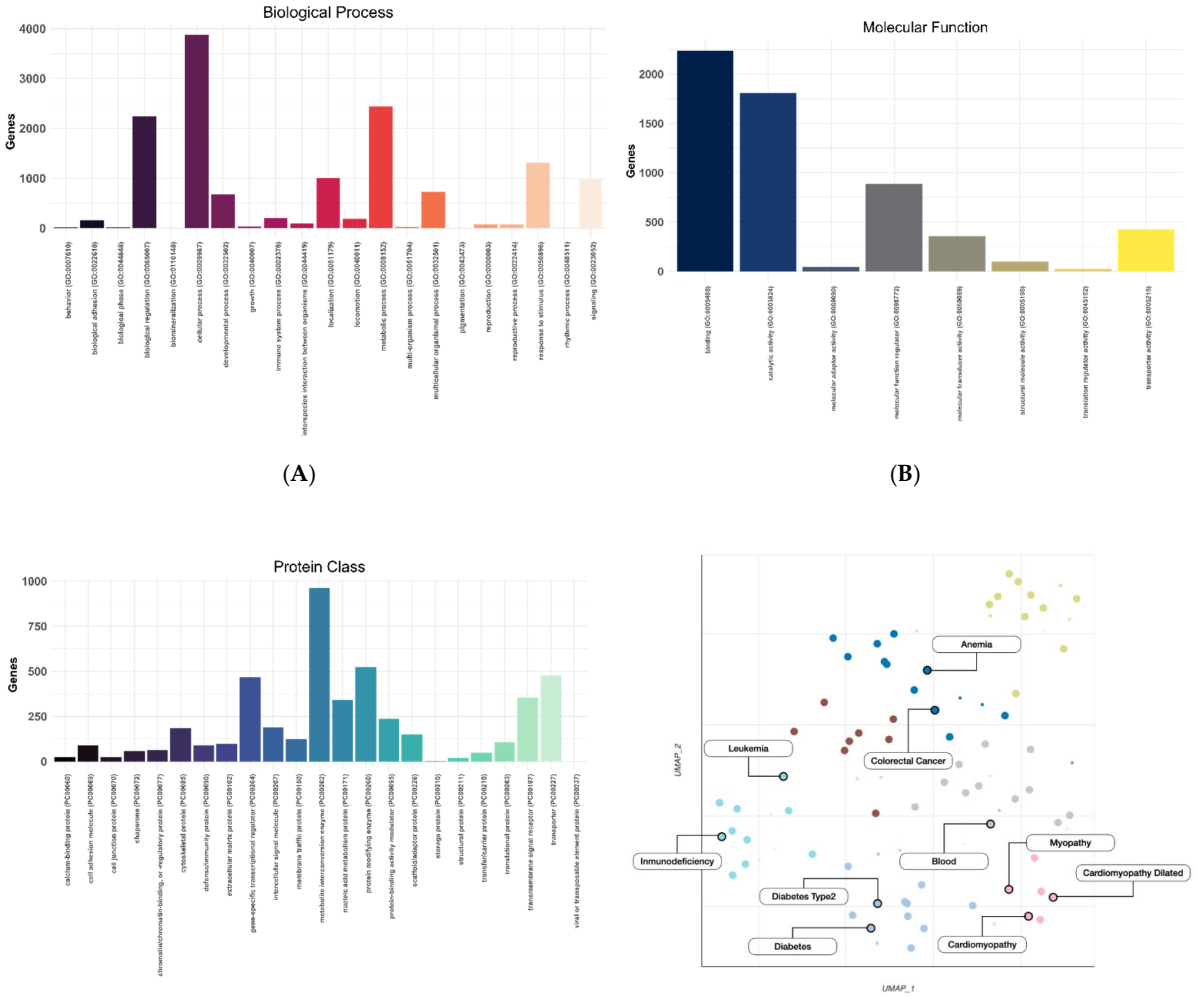


Figure 2. TSOex gene panel description: (**A**) Gene Ontology biological process (GO_BP) terms; (**B**) molecular function (GO_MF) terms; (**C**) protein class terms; (**D**) OMIM disease UMAP clustering. A, B, and C were determined using PANTHER GO-Slim gene lists. Fisher's exact test was performed using an adjusted p-value < 0.05, calculated by the Benjamini–Hochberg method. Seven OMIM disease clusters were detected using EnrichR. From the 71 significantly enriched terms (q-value < 0.05; Supplementary Table S1), the 10 most significant are labelled in panel D (adjusted p-values were calculated using the Benjamini–Hochberg method).

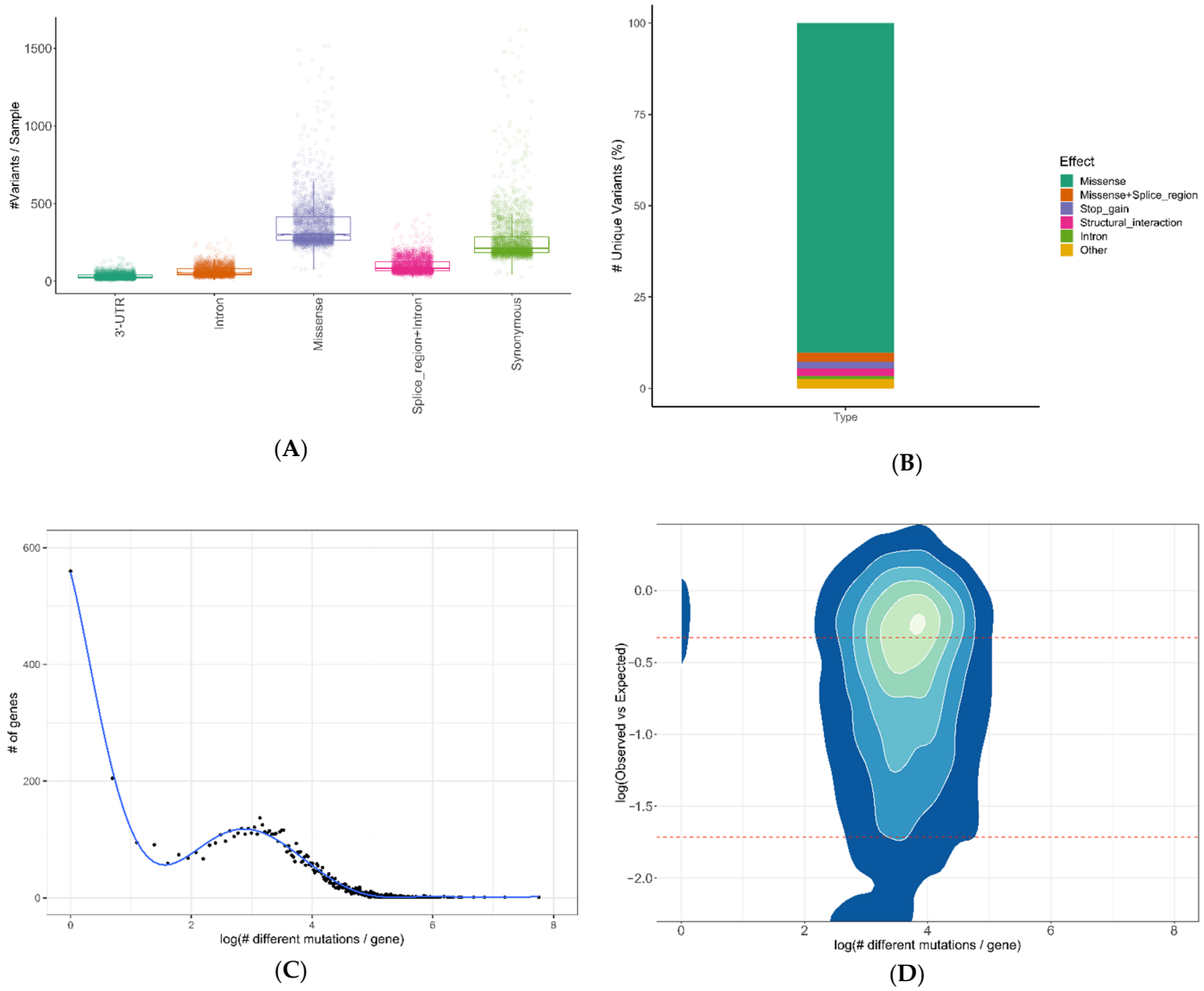
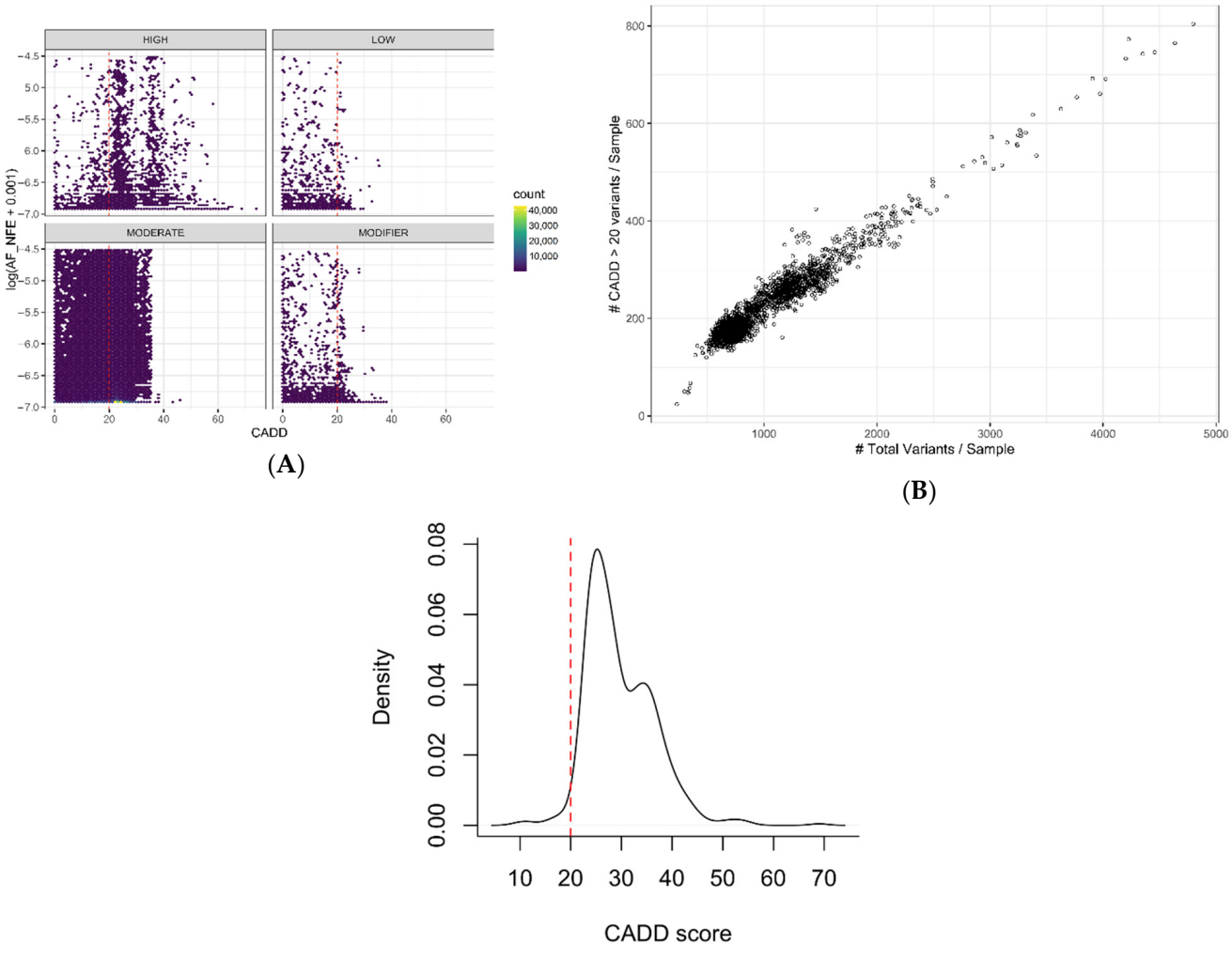


Figure 3. Genomic variability: (**A**) Number of variants detected per sample, classified by the five most abundant mutation classes. (**B**) Percentage of unique mutations detected in all 2474 patients, by class. (**C**) Genes sorted by the number of unique mutations detected (log-scaled). (**D**) Density map of gene constraints determined by gnomAD by the number of unique mutations detected (log-scaled). The red lines represent the first and forth quartiles for C and D. 3′UTR (variant in 3′ untranslated Region); intron (variant in non-coding region); missense (non-synonymous variant in coding region); splice_region + intron (splice-site variant in non-coding region); synonymous (variant in coding region that produces the same amino acid); stop-gain (variant that causes a stop codon); structural interaction (interaction loci that are likely to be supporting the protein structure).

2.2. Variant Conservation Score

Regions or individual nucleotide positions with low evolutionary variability might indicate negative selection due to functional constraints. The absence of allelic variants and a high conservation index—measured using the CADD (Combined Annotation-Dependent Depletion) index—in such regions could be an indicator of clinical importance when a variant is detected. We classified variants using the SnpEff impact categories (defined in Supplementary Table S2), the CADD score, and allele frequency (Figure 4A and Table 1). We consider CADD scores over 20 as having a high probability of being in front of a pathogenic variant. We found the correlation between HIGH (i.e., variants with a disruptive impact on the protein) mutations and CADD scores > 20 (88% within its group). Meanwhile, MODERATE (i.e., variants with a non-disruptive impact on the protein) mutations have

58% high-CADD-score variants, while LOW (i.e., variants unlikely to change protein behavior) and MODIFIER (i.e., variants with no evidence of impact) categories only reach approximately 11.5%. Most mutations have an allele frequency close to zero, indicating that putative deleterious mutations tend to be extremely rare (frequencies < 0.002 are the most common in all categories). It is worth mentioning that a high-concentration area can be observed in the MODERATE mutations, with a low allele frequency and a CADD score just above 20. When analyzing individual samples, there is a linear correlation (r = 0.97) between the total number of variants and the ones with a CADD score > 20 (Figure 4B). Approximately 20% of the mutations detected have a high CADD score independent of the sample's total mutations. This correlation might be due to the types of genes included in the TSOex gene panel. Thus, we would not expect to observe this same pattern as more regions of the genome are analyzed.



(**A**)

(**B**)

(**C**)

Figure 4. Variant conservation score: (**A**) Hexagonal heatmap of variant frequency related to the CADD score, grouped by the SnpEff impact classes (Supplementary Table S2). (**B**) Scatterplot of variants with high CADD scores (>20) related to total variants per sample. (**C**) Distribution of CADD scores of detected causative variants (n = 807). The red lines represent a CADD score of 20. AF_NFE stands for allele frequency_non-Finnish European population (gnomAD).

Table 1. Variants conservation score according to its impact class.

IMPACT	Total Number	CADD > 20	%CADD > 20
HIGH	36,753	32,207	87.63
LOW	7010	800	11.41
MODERATE	884,264	513,406	58.06
MODIFIER	13,027	1536	11.79
	941,054	547,949	58.23

2.3. Variant Classification

From the 2474 samples analyzed, approximately 800 (30%) had a single-nucleotide variant (SNV) that could be associated with the patient's pathophysiological process. According to our own data, diagnostic efficiency using TSOex reaches approximately 50%, considering other genomic variations such as indels or copy-number variants (CNVs). Almost all informed variants (98%) have a CADD score > 20 (Figure 4C), and are categorized as HIGH or MODERATE (43% and 57%, respectively). MODERATE variants have an elevated proportion of VUSs, taking into account the elevated CADD scores (Tables 1 and 2). Using the American College of Medical Genetics and Genomics (ACMG) guidelines [11], variants can also be classified into five main groups, namely, pathogenic, likely pathogenic, VUS, likely benign, and benign. Taking one sample as an example, variants were assigned to one of these five categories using the VarSome software, which calculates variant impact using the ACMG classification guidelines (Table 2) [12]. Most variants (80%) were categorized as benign or likely benign. A small fraction (<0.4%) were pathogenic or likely pathogenic. A deep phenotypic characterization is crucial to determine whether these pathogenic variants are relevant to a patient's disease. A thorough study of the overlap between the described pathophysiological alterations of the mutated gene and the patient's phenotype will lead to a causative informed variant or an unresolved study. Approximately 20% of mutations are VUSs in a randomly selected sample. After removing the benign/likely benign and pathogenic/likely pathogenic variants, a huge amount of uncertainty remains to be analyzed. As expected, most of these VUSs belong to the MODIFIER category, as their relevance in gene function or regulation is not well established. Surprisingly, a significant proportion of VUSs belong to the LOW variant class, highlighting the difficulty in reaching a conclusive diagnosis when taking only genetic data into account.

Table 2. Variant classification in one random sample.

IMPACT	Pathogenic	Likely Pathogenic	VUS [1]	Likely Benign	Benign
HIGH	11	11	65	14	119
LOW	-	4	118	388	1284
MODERATE	-	7	186	284	722
MODIFIER	-	-	1368	476	3842
	11	22	1737	1162	5967

[1]: Variant of uncertain significance.

2.4. Expression Variability

Gene expression data are gradually being incorporated into the diagnostic process to detect certain types of variations. Specifically, variability affecting alternative splicing, allele-specific outliers, and expression outliers are of great interest. The relevance of a genomic mutation with weak pathogenic evidence can be increased by its association with an aberrant gene expression profile. Transcriptomic data can be highly relevant, especially when dealing with VUS mutations. Unlike genomic data, expression data are

tissue-specific. Thus, genomic variability can be studied by considering the affected target tissue. Most studies trying to address gene expression alterations use blood or fibroblasts as surrogate tissues. The main reason to do so is that these sample types are usually much easier to obtain than the target tissue. Gene expression is extremely diverse among tissues. Clinical diagnosis must exploit transcriptomic data resources such as those generated by international consortia such as GTEx or Cell Atlas. As an example, in Figure 5A we show the expression of the 50 most expressed genes in the brain cortex distributed throughout the 54 tissues collected by the GTEx consortium. These genes are homogeneously expressed in all brain tissues, but differ greatly in others. We want to draw special attention to the difference between whole blood and brain tissues. Thus, when analyzing genomic data and inferring the putative effect of a variant, it is essential to focus on the expression data of the tissue of interest, if available. Moreover, transcript usage might also be crucial when analyzing genetic variability. Most protein-coding genes have more than one transcript (Figure 5B), with 11 being the median. *KCNMA1* has the highest number of transcripts, namely, 92. Not all transcripts are expressed equally in all tissues. Generally, one isoform is predominant, while the rest might be important in a specific tissue or in a specific stage in the cellular/organ genesis.

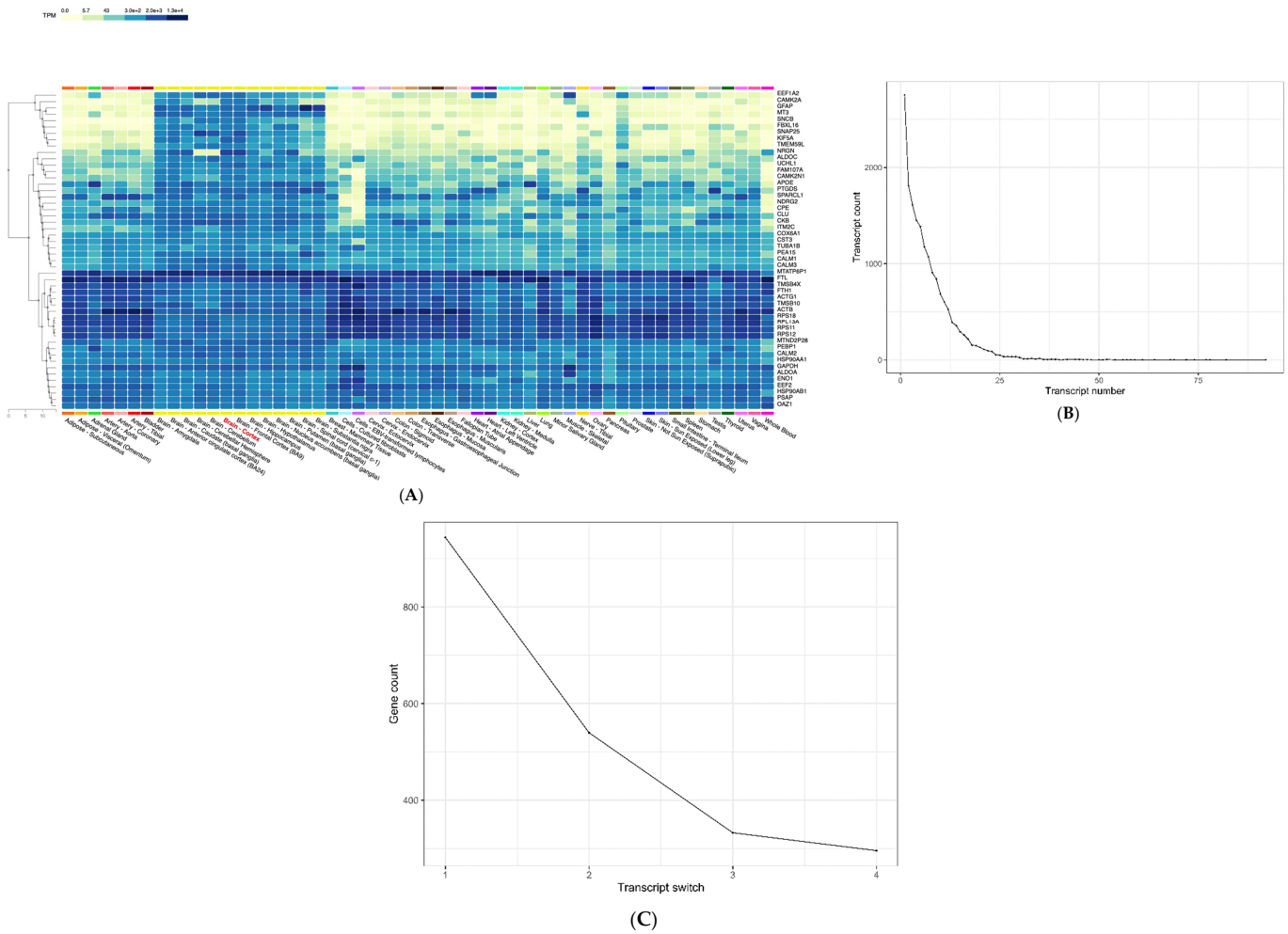
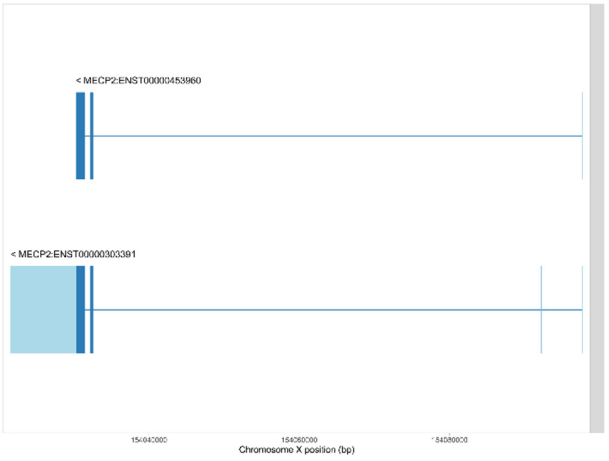


Figure 5. Transcription variability: (**A**) Tissue expression profiles of the 50 most expressed genes in the brain cortex from GTEx. (**B**) Number of protein-coding transcripts per gene using GTEx tissue data. (**C**) Rank 1 transcript switch in 1, 2, 3, or 4 out of 54 tissue datasets, using the TREGT database. Expression units are transcripts per million (TPM).

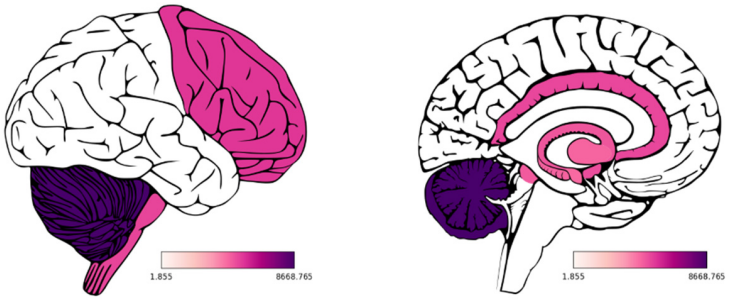
Transcript switch is not a rare event. From the 54 tissues from GTEx, the switch from the predominant isoform to another in one tissue happens in approximately 1000 genes

(Figure 5C). Tung et al. [13] analyzed up to four tissue switches, highlighting the relevance of this phenomenon. *MECP2* has two main transcripts (Figure 6A, Supplementary Table S3), and it is shown as an example of a gene with a high-order transcript switch, with 14 tissue switches (Figure 6B). The major cause of Rett syndrome (RTT) is mutations in the *MECP2* gene. RTT is a neurodevelopmental disorder, and transcript switch and differential brain region expression of *MECP2* are crucial to study mutational profiles (Figure 6C). The same *MECP2* isoform is expressed in fibroblasts and the brain, but a different one is expressed in the blood. Integration of genomic and transcriptomic data facilitates variant categorization and prioritization for enhanced diagnosis and clinical decision making.



(**A**)

(**B**)



(**C**)

Figure 6. MECP2 transcript depiction: (**A**) MECP2 schematic diagram of the two major isoforms. (**B**) Transcript switch between the different MECP2 isoforms. (**C**) MECP2 gene expression in different brain areas. In A, UTR sequences are represented in light blue. In total, 14 rank 1 transcript switches are detected among the 54 tissues studied in the TREGT database. The GTEx brain expression profile (TPMs) is represented in C using the cerebroViz R package; cerebroViz output for exterior (left) and sagittal (right) views.

2.5. Protein Variability

Protein abundance is not a replica of gene expression, and deviations at the protein level might not be detectable at the transcript level. Figure 7A shows protein quantification (high, medium, and low) for the same highly expressed genes depicted in Figure 5A (brain cortex). We obtained the data from the Human Protein Atlas. We removed genes for which there was no detectable protein in more than half of the tissues from the plot. There are significant differences between the different cell types within the cortex region, highlighting distinct cellular composition. Protein distribution contrasts with transcript

expression. Interestingly, the genes *ACTG1* and *FTH1* have much lower protein levels than their gene expression suggests, while *TUBA1A* and *PSAP* have high protein abundance, as expected.

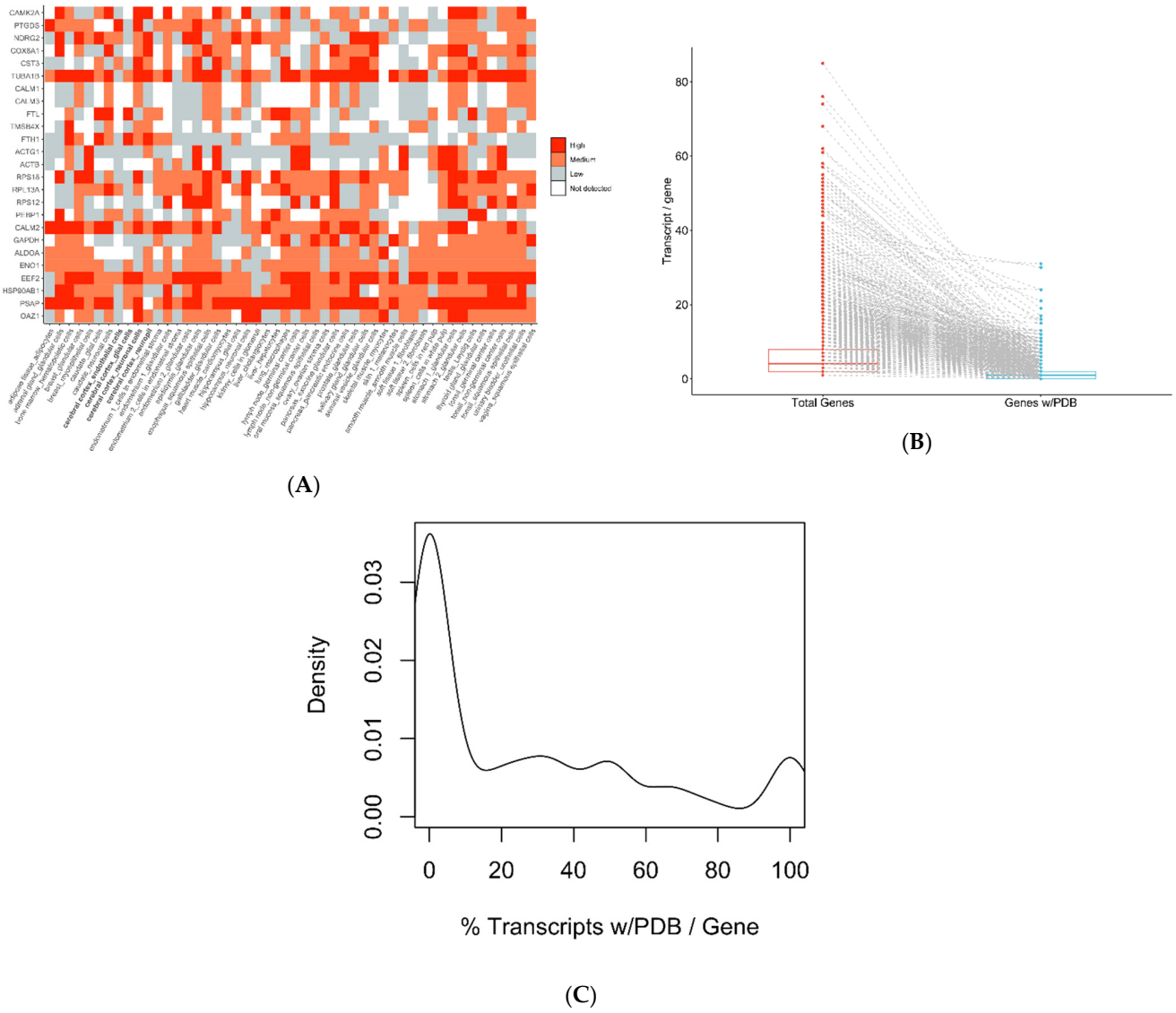


Figure 7. Protein detection and available PDB structures: (**A**) Tissue protein profiles of 25 of the 50 most expressed genes in the brain cortex from GTEx. (**B**) Gene transcripts from the TSOex panel with and without PDB IDs. (**C**) Transcript abundance with PDB IDs. In Figure 7A, proteins with no available data in >50% of the analyzed tissues from the Protein Cell Atlas were removed. In Figure 7B, the connecting dotted lines link the numbers of all transcripts from one gene with the amounts of these transcripts with PBD IDs.

2.6. Protein Structure

Secondary and tertiary protein structure can provide reliable insights into whether a mutation could be associated with disease. Studying protein structure and protein–protein interactions is another source of information to determine whether a mutation might alter the protein's function. The number of transcripts from the TSOex gene panel with known structures collected from the Protein Data Bank (PDB) is markedly smaller compared with the remaining transcripts to be characterized (Figure 7B). Most genes—approximately 50%—do not have a single transcript with a PDB structure, while only 10% have been fully resolved (Figure 7C).

3. Discussion

The incorporation of NGS into clinical diagnostics has uncovered imminent needs in terms of clarifying the huge amount of uncertainty that arises from large-scale sequencing studies. Aside from disease-specific strategies, which allow more detailed and complete designs to reach a molecular diagnosis [14–16], generic solutions when confronting any kind of rare disease are difficult to achieve. Predicting the clinical significance of a change at the DNA level is a challenging task due to the myriad of structures/processes that could be affected. In this work, we quantified the number of VUSs in the TSOex gene panel in a single clinical cohort—approximately 20% of the total detected variants, derived from a standard NGS analysis. Although conservation scores are relevant, most pathogenic variants fall into the high CADD scores (Figure 4C), which we showed to be less specific than desired, with a total of 20% of detected variants having high values (Figure 4B). Moreover, we highlighted the importance of taking into account tissue-specific transcriptomic and proteomic data to analyze the phenotypic and clinical outcomes in a more precise manner. The obvious path to elucidate the classification/pathogenicity of a genomic variant is the combination of multiple sources of information. To understand the implications of a mutation in a biological system, it is crucial to grasp the disruption caused in a highly dimensional structure such as a human being [17]. Thus, to capacitate the diagnostic process with all of the available tools and resources, reducing the variant effect uncertainty, we propose a data-aggregated tissue-specific model (Figure 8).

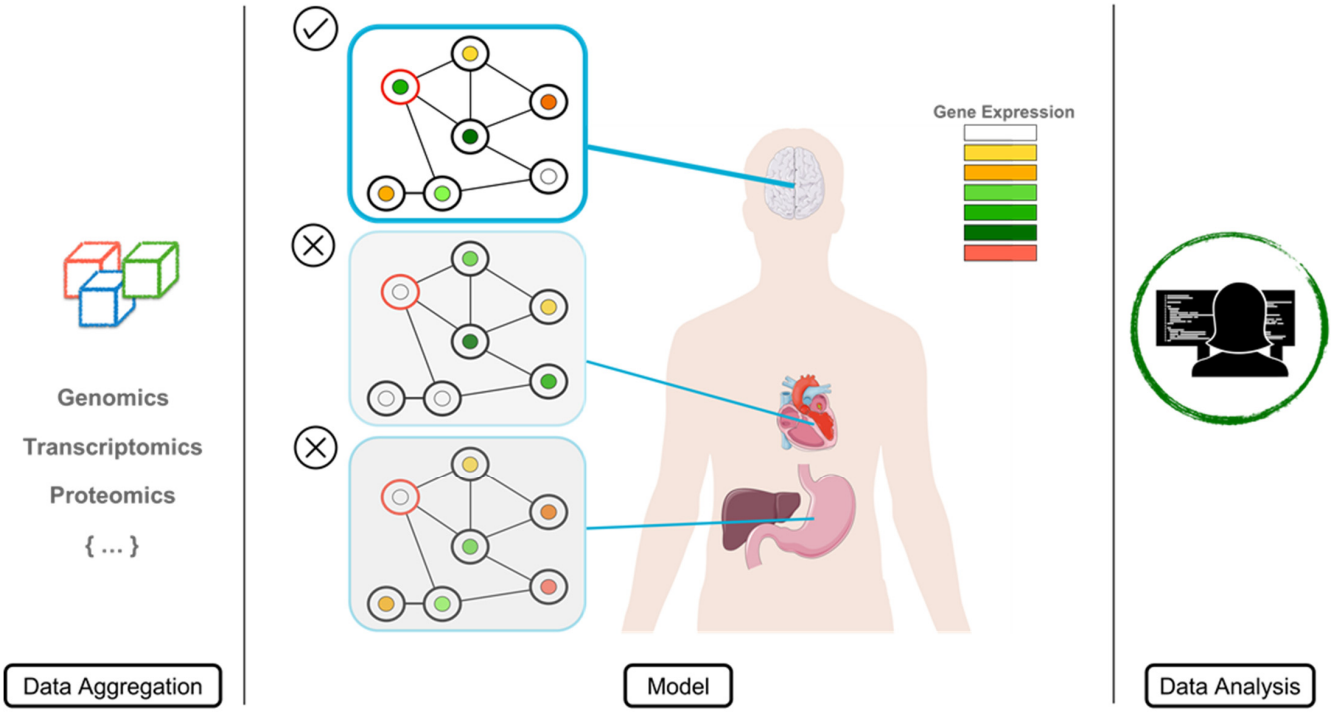


Figure 8. Multi-omics integrative analysis model.

The first step for a targeted strategy as proposed here is accurate phenotyping of the patient's symptoms and/or congenital malformations. Depending on the suspected disease, the search for the causative variant(s) will focus on a specific tissue or cell type [18]. Highly tissue-specific manifestations of genetic diseases are due to the deregulation of a functional subnetwork of genes (disease module), rather than a single gene [19]. The overall module is responsible for the tissue-specific clinical manifestation. Consequently, distinct etiological disease origins can converge in similar symptoms, leading to phenotypic overlap [20]—a many-to-one relationship. Thus, a precise clinical description may allow clinicians to hypothesize that the cause of the patient's condition is due to a specific set

of genes, narrowing down the genetic analysis to a reduced subset of putative variants. Characterizing the tissue-specific interactome is critical to find the subsets of deregulated modules and their components. Intuitively, the composition of these modules is determined by the gene–phenotype association [21–23].

Two main data components are required to build an accurate tissue-specific interactome: gene expression, and protein structure and interactions.

When focusing on a specific disease/tissue, it is important to create a context-specific interactome. The tissue-specific protein interaction landscape must be established to determine the baseline from which a perturbation might lead to the patient's disease [24,25]. Most protein interaction databases have been compiled from yeast two-hybrid (Y2H) or tandem affinity purification coupled with mass spectrometry (TAP–MS) experiments, without tissue-specific information [26]. To overcome this limitation, transcriptomics has proven to be a crucial tool to prune non-relevant interactions between proteins that in some cases might be associated, but probably not in all scenarios. GTEx and the Human Cell Atlas [27] are two commonly used sources of precise tissue- and cell-type-specific transcriptomic data, respectively, that can be combined with protein interaction data [28]. Co-expression matrices and/or gene regulatory networks can shape the tissue-specific interaction networks [29,30]. Moreover, the combination of gene expression and protein–protein interaction networks might identify gene modules that correlate with disease deregulation. When constructing the different interactomes, isoform differential functionality [31] and expression must be taken into account, as shown in Figure 6. In this regard, another issue arises: depending on the source of information, the coordinates of isoform boundaries might be slightly different. Thus, the nucleotide composition of certain isoforms is questioned. To reduce this uncertainty, a collaborative project called Matched Annotation from NCBI and EMBL-EBI (MANE) has the main objective of reporting a consensus among both reference sets [32]. Still, differences have to be computed when analyzing protein structure and interactions.

Once the main tissue-specific interactome components have been related, omics data can be combined to associate genomic variability with cell/tissue system disruption. Publicly available data (as mentioned previously) or the patient's own data can be utilized. To achieve a more accurate analysis, data-intensive precise medicine must be performed, and the patient's multi-omics data should be collected routinely [33]. Currently, this scenario in daily medical routine is unthinkable, but it should become available as omics technologies become more cost-effective. If transcriptomic and/or proteomic data from the patient are available, even from a surrogate tissue, several complementary analyses can be performed if at least some of the genes contained in the disease module are expressed. From the transcriptomics point of view, aberrant events such as expression and splicing outliers or allelic imbalance can be detected for a more individualized analysis [34,35]. Due to the fact that correlation between mRNA and protein abundance is not always detected, as shown in Figures 5A and 6A, the patient's transcriptomic data alone might be insufficient to identify the outcome of a genomic mutation. For each tissue there is a balance between both molecules. Thus, disruption of this equilibrium might indicate the source of variability that leads to a pathogenic consequence [36,37].

The presence and abundance of a protein are not the only important factors in deciphering the complexity of the interactome within a tissue; the protein's three-dimensional (3D) structure/domains and protein–protein interactions are also essential. For decades, structural biologists have studied how to predict the most stable protein folding using methods such as X-ray crystallography, nuclear magnetic resonance spectroscopy (NMR), and cryogenic electron microscopy (cryo-EM). The study of all human proteins using these methods is not possible due to time and cost constraints. Nevertheless, it is imperative to be able to study all proteins capable of triggering a genetic disease, and to study all variants detected by NGS, assigning each of them with potential pathogenicity [38]. Just recently, an algorithm using artificial intelligence, AlphaFold, has shed some light on the prediction of a protein's correct conformation [39]. The use of multi-sequence analysis

provides the possibility to transform a nucleotide sequence into a physiologically stable 3D protein structure. A huge leap forward has been accomplished by the use of deep learning algorithms. This strategy is not only limited to monomeric proteins—the same concept behind AlphaFold is being applied to multimers [40]. Furthermore, in the quest to model tissue-specific interactomes, allosteric sites are an important attribute when studying protein conformation. To map long-range communications between protein regions, a global scan of all proteins must be accomplished to ensure that variability in those sites is accountable for the putative interaction deregulation [41].

Building a specialized protein–protein interaction network (PPIN) depends heavily on prior knowledge. In addition to the protein interaction databases that base part of their protein relationship on previously mentioned techniques such as Y2H [42–44], predicted interactions can be established using several machine learning techniques. Some methods predict protein interactions using amino acid sequences and support-vector machine (SVM) algorithms with k-nearest neighbor with local description, conventional autocovariance, or deep neural networks with amphiphilic pseudo amino acid composition [45–47]. Other approaches use a genomic sequence evolutionary perspective or combination with a learning algorithm [48–50]. More recently, due to the fact that two proteins are more likely to interact if they share a common biological process or are present in the same subcellular compartment, semantic similarity has been used to predict PPINs [51]. Moreover, in specific disease scenarios it is fundamental to define the protein's subcellular localization [52], and combining Gene Ontology with SVM to predict protein interactions might be fundamental to elucidating the proper interactome. New technologies such as PROPER-seq that map protein interactions at a massive scale can help create accurate and specialized PPINs [53]. Importantly, cohesion and separation indices, as well as topological features (i.e., centrality, clustering, or node degrees), are relevant to define interactions between proteins in a PPIN [54].

Although the tissue-specific model presented in this study has been introduced in the context of rare Mendelian disease diagnostics, it is worth mentioning that it could also be applied as a more general model. Two prominent examples are polygenic and late-onset diseases. The former comprise related variants as candidate genes that compose disease modules. The combined effect of multiple mutations within the same subnetwork can be characterized. Polygenic risk scores can also be used to weight the interactions between the components of the interactome. Using this kind of approach, new disease-related genes can be discovered, and the specific relevance of each of their members can be measured [55]. The latter relates to ageing and dynamic evolution through tissue development. While at early stages of life the differentiation process requires most of the organism's energy, during ageing there is a reverse effect, leading to loss of tissue and cellular identity [56,57]. The interactome's trajectories can be shaped, and deviations over time can be modelled in a pathophysiological context [58].

Despite recent advancements in the omics realm, several limitations still prevent full characterization of variant pathogenicity. We have presented the difficulties in determining the effects that SNVs can trigger, but other sources of genomic variability might present the same challenges. Nevertheless, mutations at the gene regulatory level (e.g., sncRNAs, promoters, epigenetic signatures, or enhancers), dynamic expansions, or mutations at the genomic level (e.g., structural variants or genomic architecture) can also benefit from an interactome model [59,60]. We believe that with the introduction of new high-throughput technology—particularly in the proteomics area—and integrative algorithms to combine multidimensional data, variant effect uncertainty will be greatly reduced with a tissue-specific model.

4. Materials and Methods

4.1. Samples

DNA from 2474 pediatric patients with a rare disease was extracted from blood sampled over a 2-year period (June 2018–June 2020) at the Sant Joan de Déu Children's Hospital.

Samples were processed to capture the regions designed on the TruSight One expanded Illumina commercial gene panel (Illumina Inc., San Diego, CA, USA), which includes the coding regions and flanking intronic regions from approximately 6700 genes with known clinical phenotype association, following the protocol instructions and sequenced using a NextSeq 500 instrument (Illumina Inc., San Diego, CA, USA).

4.2. Variant Calling

Approximately 6700 genes from the TruSight One expanded clinical exome (Illumina Inc., San Diego, CA, USA) were analyzed per sample. Briefly, FASTQ files were generated using a NextSeq 500 sequencer (Illumina Inc. FastQC v0.11.5 software was used to evaluate read/base quality [61]). Adaptors and low-quality bases (Phred score < 20) were removed using cutadapt software [62]. Reads were aligned to the reference genome HG19 using BWA-MEM [63], and variant calling was performed using GATK 3.7 [64], DeepVariant v0.10.0 [65], and Octopus v0.6.3 [66]. In 807 samples, at least one pathogenic mutation was identified and reported. CNVs were not analyzed in this study.

4.3. Annotation

Mutations with gnomAD v 2.1.1 [67] European non-Finnish frequencies > 0.01 were removed from downstream analysis. CADDv1.3 and the observed versus expected ratio from gnomAD were annotated using SnpEff to determine the impacts specific mutations might have on the protein function. Plots were generated using R language with the ggplot package.

4.4. Gene/Disease Classification

PANTHER 16.0 was used to group TSOex genes in different categories [68]. OMIM disease enrichment analysis was performed using EnrichR [69]. Clusters were computed using the Leiden algorithm. Disease terms were plotted on the first two UMAP dimensions.

4.5. Databases

Transcriptomics data were collected from the GTEx web portal [70] on 12 October 2021. Brain *MECP2* expression (TPMs) from GTEx was visualized using the R package cerebroViz [71]. Transcript switch and quantification were obtained from Top-Ranked Transcript Isoforms in Human Protein-Coding Genes (TREGT) [72]. Transcript PDB IDs were extracted from ENSEMBL BioMart on 12 October 2021 [73]. Protein quantification from specific tissue cell types was retrieved from the Human Protein Atlas [74].

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/ijms23137176/s1.

Author Contributions: G.F., D.Y., F.P. and J.A. conceived the hypothesis and study design. D.Y. and J.A. performed data acquisition and variant interpretation. G.F. analyzed the data. G.F., D.Y., F.P. and J.A. drafted the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by a grant from the Spanish Ministry of Health (Instituto de Salud Carlos III/FEDER, PI20/00389) and Marató TV3-2020 (Fundació La Marató TV3, 202040-30).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to thank the Hospital Sant Joan de Déu's Genetic and Molecular Medicine department for making this study possible.

Conflicts of Interest: The authors declare no competing interests. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Lander, E.S.; Linton, L.M.; Birren, B.; Nusbaum, C.; Zody, M.C.; Baldwin, J.; Devon, K.; Dewar, K.; Doyle, M.; FitzHugh, W.; et al. Initial sequencing and analysis of the human genome. *Nature* **2001**, *409*, 860–921. [CrossRef] [PubMed]
2. Orphanet. Available online: https://www.orpha.net/consor/cgi-bin/index.php (accessed on 15 January 2022).
3. Philippakis, A.A.; Azzariti, D.R.; Beltran, S.; Brookes, A.J.; Brownstein, C.A.; Brudno, M.; Brunner, H.G.; Buske, O.J.; Carey, K.; Doll, C.; et al. The Matchmaker Exchange: A Platform for Rare Disease Gene Discovery. *Hum. Mutat.* **2015**, *36*, 915–921. [CrossRef] [PubMed]
4. Orphanet Database. Available online: www.orphadata.org (accessed on 15 January 2022).
5. Ng, S.B.; Turner, E.; Robertson, P.D.; Flygare, S.D.; Bigham, A.W.; Lee, C.; Shaffer, T.; Wong, M.; Bhattacharjee, A.; Eichler, E.E.; et al. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **2009**, *461*, 272–276. [CrossRef] [PubMed]
6. Bamshad, M.J.; Nickerson, D.A.; Chong, J.X. Mendelian Gene Discovery: Fast and Furious with No End in Sight. *Am. J. Hum. Genet.* **2019**, *105*, 448–455. [CrossRef]
7. Durmaz, A.A.; Karaca, E.; Demkow, U.; Toruner, G.; Schoumans, J.; Cogulu, O. Evolution of Genetic Techniques: Past, Present, and Beyond. *BioMed Res. Int.* **2015**, *2015*, 461524. [CrossRef]
8. Yubero, D.; Brandi, N.; Ormazabal, A.; García-Cazorla, A.; Pérez-Dueñas, B.; Campistol, J.; Ribes, A.; Palau, F.; Artuch, R.; Armstrong, J.; et al. Targeted Next Generation Sequencing in Patients with Inborn Errors of Metabolism. *PLoS ONE* **2016**, *11*, e0156359. [CrossRef]
9. Schlüter, A.; Rodríguez-Palmero, A.; Verdura, E.; Vélez-Santamaría, V.; Ruiz, M.; Fourcade, S.; Planas-Serra, L.; Martínez, J.J.; Guilera, C.; Girós, M.; et al. Diagnosis of Genetic White Matter Disorders by Singleton Whole-Exome and Genome Sequencing Using Interactome-Driven Prioritization. *Neurology* **2022**, *98*, e912–e923. [CrossRef]
10. Boycott, K.M.; Hartley, T.; Biesecker, L.G.; Gibbs, R.A.; Innes, A.M.; Riess, O.; Belmont, J.; Dunwoodie, S.L.; Jojic, N.; Lassmann, T.; et al. A Diagnosis for All Rare Genetic Diseases: The Horizon and the Next Frontiers. *Cell* **2019**, *177*, 32–37. [CrossRef]
11. Richards, S.; Aziz, N.; Bale, S.; Bick, D.; Das, S.; Gastier-Foster, J.; Grody, W.W.; Hegde, M.; Lyon, E.; Spector, E.; et al. Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **2015**, *17*, 405–423. [CrossRef]
12. Varsome, The human Genomics Community. Available online: https://varsome.com (accessed on 1 November 2021).
13. Tung, K.-F.; Pan, C.-Y.; Chen, C.-H.; Lin, W.-C. Top-ranked expressed gene transcripts of human protein-coding genes investigated with GTEx dataset. *Sci. Rep.* **2020**, *10*, 16245. [CrossRef]
14. Togi, S.; Ura, H.; Niida, Y. Application of Combined Long Amplicon Sequencing (CoLAS) for Genetic Analysis of Neurofibromatosis Type 1: A Pilot Study. *Curr. Issues Mol. Biol.* **2021**, *43*, 782–801. [CrossRef] [PubMed]
15. Bury, A.G.; Robertson, F.M.; Pyle, A.; Payne, B.A.I.; Hudson, G. The Isolation and Deep Sequencing of Mitochondrial DNA. *Methods Mol. Biol.* **2021**, *2277*, 433–447. [CrossRef] [PubMed]
16. Sorrentino, E.; Albion, E.; Modena, C.; Daja, M.; Cecchin, S.; Paolacci, S.; Miertus, J.; Bertelli, M.; Maltese, P.E.; Chiurazzi, P.; et al. PacMAGI: A pipeline including accurate indel detection for the analysis of PacBio sequencing data applied to RPE65. *Gene* **2022**, *832*, 146554. [CrossRef]
17. Noell, G.; Faner, R.; Agustí, A. From systems biology to P4 medicine: Applications in respiratory medicine. *Eur. Respir. Rev.* **2018**, *27*, 170110. [CrossRef]
18. Eraslan, G.; Drokhlyansky, E.; Anand, S.; Fiskin, E.; Subramanian, A.; Slyper, M.; Wang, J.; Van Wittenberghe, N.; Rouhana, J.M.; Waldman, J.; et al. Single-nucleus cross-tissue molecular reference maps toward understanding disease gene function. *Science* **2022**, *376*, eabl4290. [CrossRef] [PubMed]
19. Kitsak, M.; Sharma, A.; Menche, J.; Guney, E.; Ghiassian, S.D.; Loscalzo, J.; Barabási, A.-L. Tissue Specificity of Human Disease Module. *Sci. Rep.* **2016**, *6*, 35241. [CrossRef] [PubMed]
20. Vidal, S.; Brandi, N.; Pacheco, P.; Maynou, J.; Fernandez, G.; Xiol, C.; Pascual-Alonso, A.; Pineda, M.; Armstrong, J.; del Mar, O.M.; et al. The most recurrent monogenic disorders that overlap with the phenotype of Rett syndrome. *Eur. J. Paediatr. Neurol.* **2019**, *23*, 609–620. [CrossRef] [PubMed]
21. Köhler, S.; Gargano, M.; Matentzoglu, N.; Carmody, L.C.; Lewis-Smith, D.; Vasilevsky, N.A.; Danis, D.; Balagura, G.; Baynam, G.; Brower, A.M.; et al. The Human Phenotype Ontology in 2021. *Nucleic Acids Res.* **2021**, *49*, D1207–D1217. [CrossRef]
22. Martin, A.R.; Williams, E.; Foulger, R.E.; Leigh, S.; Daugherty, L.C.; Niblock, O.; Leong, I.U.S.; Smith, K.R.; Gerasimenko, O.; Haraldsdottir, E.; et al. PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. *Antonio Nat. Genet.* **2019**, *51*, 1560–1565. [CrossRef]
23. Martinez-Monseny, A.; Cuadras, D.; Bolasell, M.; Muchart, J.; Arjona, C.; Borregan, M.; Algrabli, A.; Montero, R.; Artuch, R.; Velázquez-Fragua, R.; et al. From gestalt to gene: Early predictive dysmorphic features of PMM2-CDG. *J. Med Genet.* **2018**, *56*, 236–245. [CrossRef]
24. Bossi, A.; Lehner, B. Tissue specificity and the human protein interaction network. *Mol. Syst. Biol.* **2009**, *5*, 260. [CrossRef] [PubMed]
25. Lopes, T.J.S.; Schaefer, M.; Shoemaker, J.; Matsuoka, Y.; Fontaine, J.; Neumann, G.; Andrade-Navarro, M.A.; Kawaoka, Y.; Kitano, H. Tissue-specific subnetworks and characteristics of publicly available human protein interaction databases. *Bioinformatics* **2011**, *27*, 2414–2421. [CrossRef]

26. Bajpai, A.K.; Davuluri, S.; Tiwary, K.; Narayanan, S.; Oguru, S.; Basavaraju, K.; Dayalan, D.; Thirumurugan, K.; Acharya, K.K. Systematic comparison of the protein-protein interaction databases from a user's perspective. *J. Biomed. Inform.* **2020**, *103*, 103380. [CrossRef] [PubMed]
27. Regev, A.; Teichmann, S.A.; Lander, E.S.; Amit, I.; Benoist, C.; Birney, E.; Bodenmiller, B.; Campbell, P.; Carninci, P.; Clatworthy, M.; et al. Science forum: The Human Cell Atlas. *eLife* **2017**, *6*, e27041. [CrossRef] [PubMed]
28. Glass, K.; Huttenhower, C.; Quackenbush, J.; Yuan, G.-C. Passing Messages between Biological Networks to Refine Predicted Interactions. *PLoS ONE* **2013**, *8*, e64832. [CrossRef]
29. Van Dam, S.; Võsa, U.; Van Der Graaf, A.; Franke, L.; De Magalhães, J.P. Gene co-expression analysis for functional classification and gene–disease predictions. *Brief. Bioinform.* **2018**, *19*, 575–592. [CrossRef]
30. Matched Annotation from NCBI and EMBL-EBI (MANE). Available online: https://www.ncbi.nlm.nih.gov/refseq/MANE/ (accessed on 15 January 2022).
31. Karlebach, G.; Carmody, L.; Sundaramurthi, J.C.; Casiraghi, E.; Hansen, P.; Reese, J.; Mungall, C.J.; Valentini, G.; Robinson, P.N. An algorithmic framework for isoform-specific functional analysis. *bioRxiv* **2022**. [CrossRef]
32. Weighill, D.; Ben Guebila, M.; Glass, K.; Quackenbush, J.; Platig, J. Predicting genotype-specific gene regulatory networks. *Genome Res.* **2022**, *32*, 524–533. [CrossRef]
33. Menyhárt, O.; Győrffy, B. Multi-omics approaches in cancer research with applications in tumor subtyping, prognosis, and diagnosis. *Comput. Struct. Biotechnol. J.* **2021**, *19*, 949–960. [CrossRef]
34. Ferraro, N.M.; Strober, B.J.; Einson, J.; Abell, N.S.; Aguet, F.; Barbeira, A.N.; Brandt, M.; Bucan, M.; Castel, S.E.; Davis, J.R.; et al. Transcriptomic signatures across human tissues identify functional rare genetic variation. *Science* **2020**, *369*, eaaz5900. [CrossRef]
35. Yépez, V.A.; Mertes, C.; Müller, M.F.; Klaproth-Andrade, D.; Wachutka, L.; Frésard, L.; Gusic, M.; Scheller, I.F.; Goldberg, P.F.; Prokisch, H.; et al. Detection of aberrant gene expression events in RNA sequencing data. *Nat. Protoc.* **2021**, *16*, 1276–1296. [CrossRef] [PubMed]
36. Kopajtich, R.; Smirnov, D.; Stenton, S.L.; Loipfinger, S.; Meng, C.; Scheller, I.F.; Freisinger, P.; Baski, R.; Berutti, R.; Behr, J.; et al. Integration of proteomics with genomics and transcriptomics increases the diagnostic rate of Mendelian disorders. *medRxiv* **2021**, 1–31. [CrossRef]
37. Du, Y.; Clair, G.C.; Al Alam, D.; Danopoulos, S.; Schnell, D.; Kitzmiller, J.A.; Misra, R.S.; Bhattacharya, S.; Warburton, D.; Mariani, T.J.; et al. Integration of transcriptomic and proteomic data identifies biological functions in cell populations from human infant lung. *Am. J. Physiol. Cell. Mol. Physiol.* **2019**, *317*, L347–L360. [CrossRef] [PubMed]
38. Kustatscher, G.; Collins, T.; Gingras, A.-C.; Guo, T.; Hermjakob, H.; Ideker, T.; Lilley, K.S.; Lundberg, E.; Marcotte, E.M.; Ralser, M.; et al. Understudied proteins: Opportunities and challenges for functional proteomics. *Nat. Methods* **2022**. Online ahead of print. [CrossRef]
39. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583–589. [CrossRef]
40. Evans, R.; O'Neill, M.; Pritzel, A.; Antropova, N.; Senior, A.; Green, T.; Žídek, A.; Bates, R.; Blackwell, S.; Yim, J.; et al. Protein complex prediction with AlphaFold-Multimer. *bioRxiv* **2021**. [CrossRef]
41. Faure, A.J.; Domingo, J.; Schmiedel, J.M.; Hidalgo-Carcedo, C.; Diss, G.; Lehner, B. Global mapping of the energetic and allosteric landscapes of protein binding domains. *bioRxiv* **2021**. [CrossRef]
42. Orchard, S.; Ammari, M.; Aranda, B.; Breuza, L.; Briganti, L.; Broackes-Carter, F.; Campbell, N.H.; Chavali, G.; Chen, C.; Del-Toro, N.; et al. The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.* **2013**, *42*, D358–D363. [CrossRef]
43. Szklarczyk, D.; Gable, A.L.; Nastou, K.C.; Lyon, D.; Kirsch, R.; Pyysalo, S.; Doncheva, N.T.; Legeay, M.; Fang, T.; Bork, P.; et al. The STRING database in 2021: Customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res.* **2020**, *49*, D605–D612. [CrossRef]
44. Fahey, M.E.; Bennett, M.J.; Mahon, C.; Jäger, S.; Pache, L.; Kumar, D.; Shapiro, A.; Rao, K.; Chanda, S.K.; Craik, C.S.; et al. GPS-Prot: A web-based visualization platform for integrating host-pathogen interaction data. *BMC Bioinform.* **2011**, *12*, 298. [CrossRef]
45. Xia, J.; Gui, J. Prediction of Protein-Protein Interactions from Protein Sequence Using Local Descriptors. *Protein Pept. Lett.* **2010**, *17*, 1085–1090. [CrossRef] [PubMed]
46. Guo, Y.; Yu, L.; Wen, Z.; Li, M. Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences. *Nucleic Acids Res.* **2008**, *36*, 3025–3030. [CrossRef] [PubMed]
47. Du, X.; Sun, S.; Hu, C.; Yao, Y.; Yan, Y.; Zhang, Y. DeepPPI: Boosting Prediction of Protein–Protein Interactions with Deep Neural Networks. *J. Chem. Inf. Model.* **2017**, *57*, 1499–1510. [CrossRef] [PubMed]
48. Tuncbag, N.; Gursoy, A.; Nussinov, R.; Keskin, O. Predicting protein-protein interactions on a proteome scale by matching evolutionary and structural similarities at interfaces using PRISM. *Nat. Protoc.* **2011**, *6*, 1341–1354. [CrossRef]
49. Zhang, L.V.; Wong, S.L.; King, O.D.; Roth, F.P. Predicting co-complexed protein pairs using genomic and proteomic data integration. *BMC Bioinform.* **2004**, *5*, 38. [CrossRef]
50. Li, F.; Zhu, F.; Ling, X.; Liu, Q. Protein Interaction Network Reconstruction through Ensemble Deep Learning with Attention Mechanism. *Front. Bioeng. Biotechnol.* **2020**, *8*, 390. [CrossRef]

51. Armean, I.M.; Lilley, K.S.; Trotter, M.W.B.; Pilkington, N.C.V.; Holden, S.B. Co-complex protein membership evaluation using Maximum Entropy on GO ontology and InterPro annotation. *Bioinformatics* **2018**, *34*, 1884–1892. [CrossRef]
52. Hooper, C.M.; Castleden, I.R.; Tanz, S.K.; Grasso, S.V.; Millar, A.H. Subcellular Proteomics as a Unified Approach of Experimental Localizations and Computed Prediction Data for Arabidopsis and Crop Plants. *Adv. Exp. Med. Biol.* **2021**, *1346*, 67–89. [CrossRef]
53. Johnson, K.L.; Qi, Z.; Yan, Z.; Wen, X.; Nguyen, T.C.; Zaleta-Rivera, K.; Chen, C.-J.; Fan, X.; Sriram, K.; Wan, X.; et al. Revealing protein-protein interactions at the transcriptome scale by sequencing. *Mol. Cell* **2021**, *81*, 4091–4103.e9. [CrossRef]
54. Ying, K.-C.; Lin, S.-W. Maximizing cohesion and separation for detecting protein functional modules in protein-protein interaction networks. *PLoS ONE* **2020**, *15*, e0240628. [CrossRef]
55. Bern, M.; King, A.; Applewhite, D.A.; Ritz, A. Network-based prediction of polygenic disease genes involved in cell motility. *BMC Bioinform.* **2019**, *20*, 313. [CrossRef] [PubMed]
56. Wang, X.; Jiang, Q.; Song, Y.; He, Z.; Zhang, H.; Song, M.; Zhang, X.; Dai, Y.; Karalay, O.; Dieterich, C.; et al. Ageing induces tissue-specific transcriptomic changes in *Caenorhabditis elegans*. *EMBO J.* **2022**, *41*, e109633. [CrossRef] [PubMed]
57. Izgi, H.; Han, D.; Isildak, U.; Huang, S.; Kocabiyik, E.; Khaitovich, P.; Somel, M.; Dönertaş, H.M. Inter-tissue convergence of gene expression during ageing suggests age-related loss of tissue and cellular identity. *eLife* **2022**, *11*, e68048. [CrossRef] [PubMed]
58. Fu, D.; He, J. DPPIN: A Biological Repository of Dynamic Protein-Protein Interaction Network Data. *arXiv* **2017**, 02168. [CrossRef]
59. Zhang, L.; Lu, Q.; Chang, C. Epigenetics in Health and Disease. *Adv. Exp. Med. Biol.* **2020**, *1253*, 3–55. [CrossRef] [PubMed]
60. Mishra, A.; Hawkins, R.D. Three-dimensional genome architecture and emerging technologies: Looping in disease. *Genome Med.* **2017**, *9*, 87. [CrossRef] [PubMed]
61. Babraham Bioinformatics. Available online: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (accessed on 15 January 2022).
62. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* **2011**, *17*, 10–12. [CrossRef]
63. Li, H.; Durbin, R. Fast and accurate short read alignment with Burrows—Wheeler transform. *Bioinformatics* **2009**, *25*, 1754–1760. [CrossRef]
64. Van der Auwera, G.A.; O'Connor, B.D. *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra*; O'Reilly Media: Sebastopol, CA, USA, 2020.
65. Cooke, D.P.; Wedge, D.C.; Lunter, G. A unified haplotype-based method for accurate and comprehensive variant calling. *Nat. Biotechnol.* **2021**, *39*, 885–892. [CrossRef]
66. Poplin, R.; Chang, P.-C.; Alexander, D.; Schwartz, S.; Colthurst, T.; Ku, A.; Newburger, D.; Dijamco, J.; Nguyen, N.; Afshar, P.T.; et al. A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* **2018**, *36*, 983–987. [CrossRef]
67. Karczewski, K.J.; Francioli, L.C.; Tiao, G.; Cummings, B.B.; Alfoldi, J.; Wang, Q.; Collins, R.L.; Laricchia, K.M.; Ganna, A.; Birnbaum, D.P.; et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **2020**, *581*, 434–443. [CrossRef] [PubMed]
68. Mi, H.; Ebert, D.; Muruganujan, A.; Mills, C.; Albou, L.-P.; Mushayamaha, T.; Thomas, P.D. PANTHER version 16: A revised family classification, tree-based classification tool, enhancer regions and extensive API. *Nucleic Acids Res.* **2020**, *49*, D394–D403. [CrossRef] [PubMed]
69. Chen, E.Y.; Tan, C.M.; Kou, Y.; Duan, Q.; Wang, Z.; Meirelles, G.V.; Clark, N.R.; Ma'Ayan, A. Enrichr: Interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinform.* **2013**, *14*, 128. [CrossRef] [PubMed]
70. Genotype-Tissue Expression (GTEx) Project. Available online: https://gtexportal.org (accessed on 15 January 2022).
71. Bahl, E.; Koomar, T.; Michaelson, J. cerebroViz: An R package for anatomical visu-alization of spatiotemporal brain data. *Bioinformatics* **2016**, *33*, 762–763. [CrossRef] [PubMed]
72. Smedley, D.; Haider, S.; Ballester, B.; Holland, R.; London, D.; Thorisson, G.; Kasprzyk, A. BioMart—biological queries made easy. *BMC Genom.* **2009**, *10*, 22. [CrossRef]
73. Uhlén, M.; Fagerberg, L.; Hallström, B.M.; Lindskog, C.; Oksvold, P.; Mardinoglu, A.; Sivertsson, Å.; Kampf, C.; Sjöstedt, E.; Asplund, A.; et al. Tissue-Based Map of the Human Proteome. *Science* **2015**, *347*, 1260419. [CrossRef]
74. The Human Protein Atlas. Available online: https://www.proteinatlas.org (accessed on 15 January 2022).